Learning of Spatial Invariances for Object-Ground Separation

Alex Schwarz, Frederik Beuth, Fred H Hamker
{alexschw,beuth,fhamker}@hrz.tu-chemnitz.de

TU Chemnitz, Professorship for Artificial Intelligence

Abstract. Whenever a biological plausible model should learn object representations in a real world scenario, the model tends to learn connections not only to features of the learned object but also to some features of the background. We present a learning algorithm that is capable of learning background-independent object representations due to temporal differences in the background. The algorithm is based on a previous one, used for learning the object representations in the visual attention model of Beuth and Hamker [2]. As the representations in this model have encoded a black background among the object, the model performed very well when localizing an object before a black-background, but only averagely when the object is placed in a real-world scene. To address such problems, we advance the learning algorithm to obtain background-independent object representations, and adapt the V1-layer to generate responses similar to their biological counterparts.

1 Introduction

Many biological plausible models of visual attention exist, but only a few have been shown with real-world scenes [1, 3, 6, 11]. It is still unsolved how objects must be represented to allow such models an operation in real world scenarios. Beuth and Hamker(2015) presented an attention model capable of operating on real-world scenes [2]. This model learned the object representations with the principle of temporal continuity [4, 10], as some authors suppose that the primate brain uses this mechanism for the development of invariant objects representations. Temporal continuity learning exploits the fact that, on a short time scale, changes in the visual input are more likely to originate from the same object, rather than from a different one. Temporal continuity learning algorithms have been typically used in simple scenes consisting of black bars, but not in a real-world task. In previous works [1, 2], the learning was already improved to operate in natural scenes by introducing a high post-synaptic threshold. Such a threshold allows to learn cells which are highly specific for their encoded stimuli, which perform well in setups with natural scenes. However, these object representations are still not encoding the object independently of the background. The reason is that the objects are learned from small image patches which contain the object along with some background. As the procedure learns an object representation from all available features in such a patch, it learns the objects along with the background. However, an object representation independent of the background would be preferable. It is currently unclear how the human brain learns background invariant object representations. Suggestions cover the usage of disparities [5] or motion [9], whereas we show here that temporal continuity alone is powerful enough.

In this article, we will firstly explain the existing model and learning algorithm [2], before outlining its disadvantages and presenting our solution. Secondly, we will show a brief overview of the implemented model mechanisms. Finally, we will show the details of the novel learning algorithm and demonstrate its performance by a miniature version of the model.

2 Model

2.1 The attention model

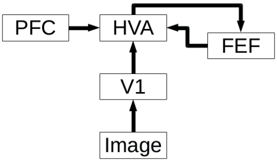

Fig. 1. Schematic layout of the model, showing the main sequence of visual areas working together

The model of Beuth and Hamker [2] consists of several layers representing the areas of the visual cortex. The first and lowest layer is the V1-layer, encoding color differences and simple orientations. The latter is constructed out of differently oriented Gabor-Filters, functions of a cosine and a Gaussian. Gabor-Filters are used because their shape fits the shape of typical receptive fields of V1-neurons [7]. Every modeled V1-neuron is then connected to a layer of HVA-neurons. These neurons of a higher visual area (HVA) are a representation of neurons in area V4 and the inferior-temporal cortex IT. In these cells an object representation is constructed out of the responses of the V1-neurons. A feature-based attention signal is then send from the prefrontal cortex (PFC) to the object-view specific cells in HVA, altering the strength of their response. The loop to the frontal eye field (FEF) will then find the correct location of the given object and propagate it back to the corresponding location in HVA.

2.2 The existing temporal-continuity learning

The construction of feature specific HVA-cells is performed by learning the individual weights from each V1-cell to each HVA-cell with a temporal-continuity learning algorithm (see eq. 1)

$$\begin{gathered}\tau_w \frac{\partial w}{\partial t} = u_t \cdot v_{t-1} - \alpha \cdot w \cdot v_{t-1}^2 \\ with : u = \left(r^{\mathrm{V1}} - \theta^{\mathrm{V1}}\right),\ v = \left(r^{\mathrm{HVA}} - \theta^{\mathrm{HVA}}\right)^+\end{gathered} \tag{1}$$

where r^{V1} and r^{HVA} are the responses of the presynaptic V1-cell and the post-synapic HVA-cell with their respective thresholds θ^{V1} and θ^{HVA}, τ_w being the learning rate and α the strength of the alpha-constraint. The alpha-constraint is an inhibitory term for the normalization and soft-limitation for the strength of weights (w) between the pre- and postsynaptic cells. The $()^+$ is a symbol for a cutoff of negative values.

2.3 The novel temporal-continuity learning algorithm

When using the temporal-continuity learning algorithm (section 2.2) to achieve background-invariant object representation, the objects are represented in the connections from V1 to HVA. The temporal continuity alone is theoretically powerful enough to learn background invariant HVA-cells on any given background. As the background changes much more often than the object, the learning rule should not learn connections from the background region, and weights only from the features of an object. Yet, we observed that inhibitory weights to the background were learned mistakenly. We found out that the reason is a difference in the learning speed for weights with positive or negative weight changes. Due to this, we introduce two normalization procedures of τ to ensure that the learning speed for both cases is balanced and that weights are generally faster decreasing towards zero (Sec. 4).

3 Implementation Details

3.1 Input Images

We use an image library, consisting of images from 100 objects in 72 different angles (COIL-100, [8]), to learn object-relevant features based on real data. To enable the learning of background-invariant cells, we need to expand the object dataset to show the objects not only on black backgrounds, but also on differently colored backgrounds, gaussian-white noise and patches from real world scenes. We systematically vary the background for each object view, using 1 black, 3 gaussian-white noise, 7 colored, and 9 random real world backgrounds (Fig. 2). We used 15 out of these 100 objects for a faster learning, thereby resulting in 21600 input images ($21600 = (1 + 3 + 7 + 9) \cdot 15 \cdot 72$).

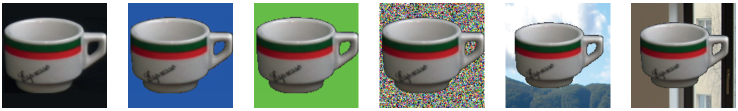

Fig. 2. Examples of the used object patches, showing one black, two colored, a noisy and two real world patches from left to right.

3.2 Normalization of the V1-cell responses

Cells of area V1 are responding equally to different lighting conditions since the retinal receptors habituate to changing intensities of light. Our model should be able to mimic this behavior. That is accomplished by the V1-cells depending on differences in contrast instead of absolute intensity values. Therefore we normalized the responses of the V1-neurons to roughly fit a range of 0 and 1.

3.3 Filter for equiluminant borders

The gabor filters in the previous model have been run on a gray-scaled version of the image. Human vision on the other hand is still able to extract features of objects and classical borders, even when receiving a colored border which is equiluminant and thus does not appear in gray-scale. To model this ability, we let the model receive input from border-detecting cells for different colors, and combine their input to an equiluminant input. As a result these border cells increased the identification performance, when added to the color based filters.

3.4 A miniature version of the learning algorithm

We included a miniature version of the learning algorithm for demonstrate the effects of several parameters easily. The miniature model consists of all features of the complete model, but does not receive any image as an input. Instead the model receives an input value of one for an active cell as if it responds to a preferred stimulus, or zero as a non-preferred input. It will then receive equally distributed random input for mimicking cells of the background. The ratio of those three cell conditions could be individually adjusted. By that we are able to analyze problems when using the model more effectively and calculate the outcome of the model when receiving an optimal input.

3.5 Changing the postsynaptic threshold θ^{HVA}

The postsynaptic threshold θ^{HVA} is typically chosen equal to the mean response of all HVA-cell neurons (population mean). In a larger population of cells, this value is very low as the HVA-cells tend to respond sparse to a given input, and thereby most of the neurons will not fire. When the threshold is too low the cells will learn not only their preferred input, but also an input with close to similar but not equal properties e.g. a wide range of orientations.

Temporal population-based θ^{HVA} Because the maximum activity of the HVA-cells is fluctuating in a wide range, one may instead use the temporal population-based mean activity as a threshold. This threshold is calculated incrementally using:

$$\theta^{\mathrm{HVA}}(t) = \theta^{\mathrm{HVA}}(t-1) + \frac{1}{t} \cdot \Big(< r^{\mathrm{HVA}}(t) > - \, \theta^{\mathrm{HVA}}(t-1) \Big) \qquad (2)$$

Constant threshold θ^{HVA} The threshold is set to a constant value, e.g. 0.5, and does not change within the learning, producing a better transferability to a complete model with more cells, due to independence from the number of V1-neurons. It works fine as long as the threshold is not higher then the overall activity.

4 Results

The novel learning algorithm normalizes the learning speed based on two components: a faster learning of weights towards zero and a balancing of the learning speed for weights with positive or negative weight changes.

4.1 Faster weight changes towards zero

When learning on real world images, the weights of cells responding to the background equilibrate at a value that is not zero, due to a difference of responses in the input dataset. Therefore we need to compensate this error by a separate scaling of weights changing towards zero in comparison to those developing away from zero.

$$\tau_w = \begin{cases} \frac{\tau_{FF}}{\zeta} & , W \rightarrow 0 \\ \tau_{FF} & , else \end{cases} \tag{3}$$

Whenever a cell changes its weight in the direction of zero, it will learn faster with a factor of the chosen parameter ζ. We have chosen a value of 2.5 after several tests. With that value, the weights to background-cells are running toward zero (see Fig. 3), resulting in a better noise-suppression as well as a good figure ground separation.

4.2 Balancing of the learning rate

When applying a presynaptic threshold for the input neurons, the response of a given input gets shifted. This was tested on a normally distributed input. Based on the height of the threshold, the learning speed should be adapted as follows to prevent a shift towards positive or negative weights.

$$\tau_{w,new} = \tau_{w,old} \cdot \tau_{pre}, \quad where : \tau_{pre} = \begin{cases} \frac{1}{1-\theta_{V1}} & , \Delta W > 0 \\ \frac{1}{\theta_{V1}} & , \Delta W < 0 \\ 1 & , \Delta W = 0 \end{cases} \tag{4}$$

These two modifications result in a lower response to non-specific V1-cells and thereby a better noise-resistance of the HVA-layer, as visible in Figure 3.

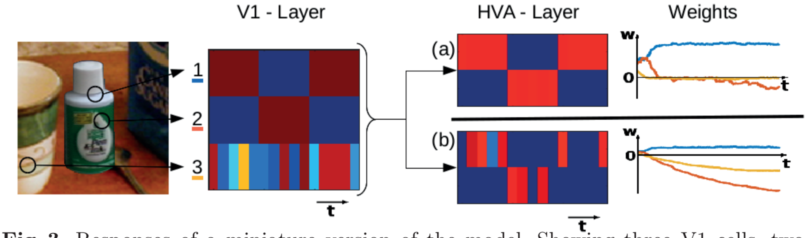

Fig. 3. Responses of a miniature version of the model. Showing three V1 cells, two HVA cells and the synaptic weights towards the first (upper) HVA cell for the complete learning time. a) With normalization of τ the HVA responses match the V1-input, b) without normalization the cells do not respond correctly because they get inhibited by the activity of the background cell(3, yellow). We used the values: $\alpha = 400, \ \tau_{FF} = 285$

5 Conclusion and future work

With the given modifications, the model is now able to learn background-invariant object representations in a miniature model. They show a good suppression of unpredictable responses of the V1-cells. In a complete model, the system still tends to learn slightly negative weights for background positions. We ascribe this problem putatively to the different number of V1 and HVA cells between both models. Several of the models parameters depend on the number of neurons, thus a finer tuning of the parameters is necessary. We are currently conducting a study solving these issues. Nevertheless, the current study shows already how the human brain may operate to achieve figure-ground separation and learn background-invariant object representations.

References

1. Antonelli, M., Gibaldi, A., Beuth, F., Duran, A.J., Canessa, A., Chessa, M., Solari, F., Del Pobil, A.P., Hamker, F., Chinellato, E., et al.: A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot. Autonomous Mental Development, IEEE Transactions on 6(4), 259–273 (2014)
2. Beuth, F., Hamker, F.H.: Attention as cognitive, holistic control of the visual system (2015)
3. Chikkerur, S., Serre, T., Tan, C., Poggio, T.: What and where: A bayesian inference theory of attention. Vision research 50(22), 2233–2247 (2010)
4. Földiák, P.: Learning Invariance from Transformation Sequences. Neural Computation 3(2), 194–200 (1991)
5. Goerick, C., Wersing, H., Mikhailova, I., Dunn, M.: Peripersonal space and object recognition for humanoids. In: Humanoid Robots, 2005 5th IEEE-RAS International Conference on. pp. 387–392. IEEE (2005)
6. Hamker, F.H.: The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. Cerebral Cortex 15(4), 431–447 (2005)
7. Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. The Journal of Physiology 148(3), 574–591 (1959)
8. Nene, S.A., Nayar, S.K., Murase, H.: "Columbia Object Image Library (COIL-100)" (1996)
9. Raudies, F., Neumann, H.: A neural model of the temporal dynamics of figure-ground segregation in motion perception. Neural Netw 23, 160–176 (2010)
10. Spratling, M.W.: Learning viewpoint invariant perceptual representations from cluttered images. IEEE 27(5), 753–761 (2005)
11. Walther, D.B., Koch, C.: Attention in hierarchical models of object recognition. Progress in brain research 165, 57–78 (2007)